Two Counterexamples to *Tokenization and the Noiseless Channel*

Marco Cognetta[1] Vilém Zouhar[2]
Sangwhan Moon[1] Naoaki Okazaki[1]

1. Tokyo Institute of Technology
2. ETH Zurich

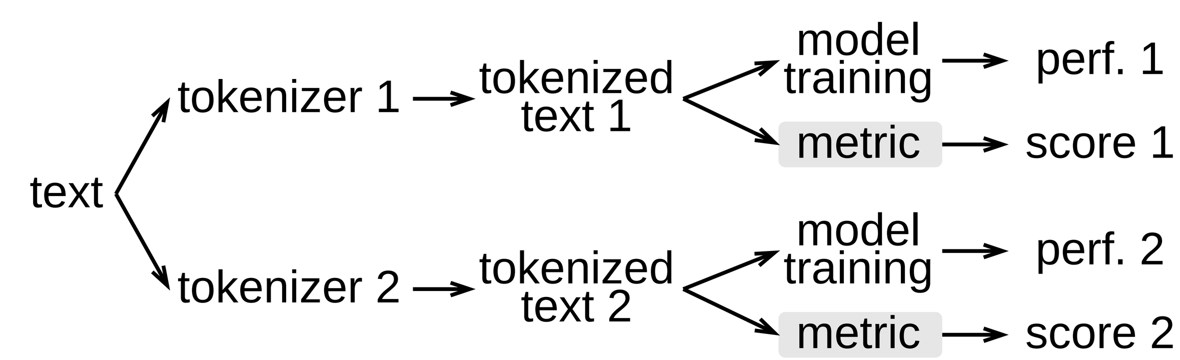
東京工業大学 Tokyo Institute of Technology

TLDR: We describe two families of BPE tokenizer variants which are counterexamples to *the Efficiency Hypothesis* and other intrinsic tokenizer metrics.

Background

- Tokenizers are often overlooked
  - They are difficult to tune
  - Surprisingly large downstream effects
- *Tokenization and the Noiseless Channel*
  - Compared intrinsic metrics
  - Found Rényi Efficiency to be the best
  - Propose *the Efficiency Hypothesis*
- We find two counter examples to it
  - Tokenizers that increase efficency but decrease BLEU

Extrinsic Train-Eval Loop



Prior Intrinsic Tokenizer Metrics

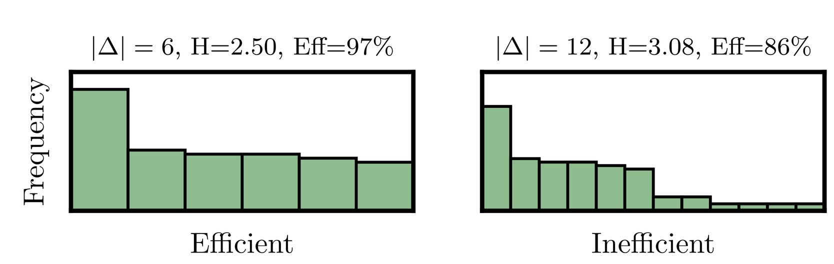
- Sequence Length
  - Average number of tokens in a sequence
  - Lower is better
- Percentile Frequency
  - Total unigram probability of $[a, b]$ percentile
  - Higher is better

Entropy Based Metrics

- Shannon Entropy
  - $H(V) = \sum_{w \in V} p(w) \log p(w)$
  - Higher is better
- Shannon Efficiency
  - $\textsc{Eff}(V) = \frac{H(V)}{\log(|V|)}$
  - Easy to compare across vocabulary sizes

Rényi Efficiency

- $H_\alpha(V) = \lim_{\alpha' \to \alpha} \frac{1}{1-\alpha'} \sum_{v \in V} \log\left(p(v)^{\alpha'}\right)$
  - $\alpha = 1$ is Shannon Entropy
  - $\alpha > 1$ encourages "flatter" distributions
- $\textsc{Eff}_\alpha(V) = \frac{H_\alpha(V)}{\log(|V|)}$



Tokenization and the Noiseless Channel

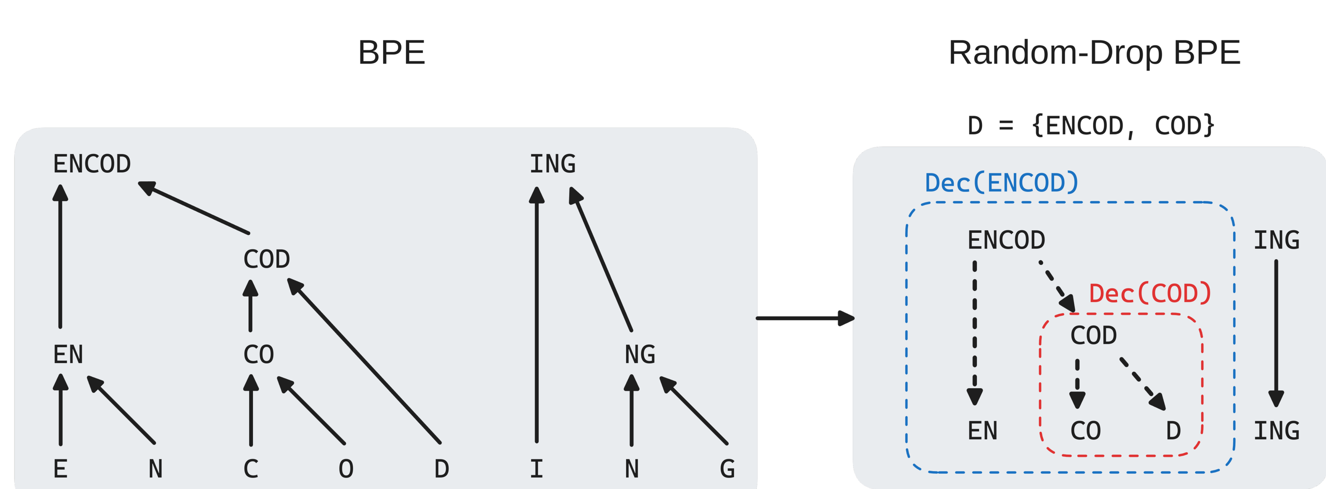
- Large scale translation task
  - Rènyi Efficiency ($\alpha = 2.7$) was best correlated to BLEU

Predictor	**Pearson**	**Spearman**	ρ^2
Sequence len.	−0.32 (=0.118)	−0.24 (=0.239)	10%
Percentile freq.	0.76 (<0.001)	0.63 (<0.001)	58%
Entropy	0.22 (=0.281)	0.12 (=0.578)	5%
Entropy eff.	0.56 (=0.004)	0.38 (=0.006)	31%
Rényi entropy	0.49 (=0.001)	0.38 (=0.006)	24%
Rényi eff.	0.78 (<0.001)	0.66 (<0.001)	61%

Table 1: Correlations between different predictors and MT performance (BLEU). The p-values for each statistic (computed using a t-test) are in parentheses.

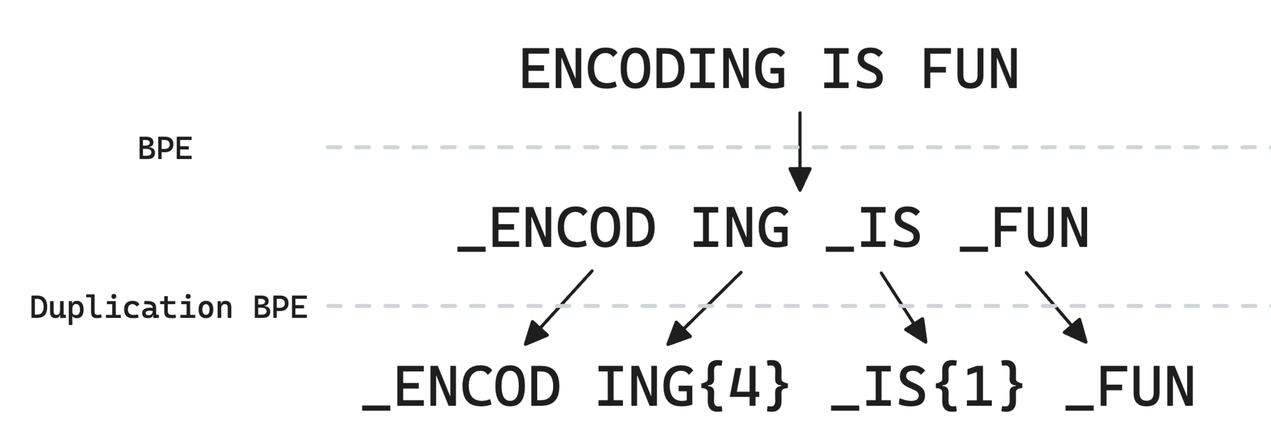
Random-Drop BPE

- Randomly mark k tokens from the top N
- After tokenization, decompose any of the marked tokens
- Provable sufficient conditions for efficiency increase



Duplicate BPE

- Create k duplicates of the top N tokens
- After tokenization, replace at random with duplicate
- Provably increases efficiency
- Intuitively decreases BLEU



Experimental Results

- DE-EN Translation Task
- Picked multiple BPE baselines
  - Varied initial vocabulary size
- Built Duplicate/Random-Drop on top of them
  - Varied N and k for each
  - Efficiency is consistently higher than baseline
  - BLEU is consistently lower than baseline

Tokenizer	N	k	Overall Eff$_\alpha$	Overall BLEU	Best Eff$_\alpha$*	Best BLEU*
Baseline (4k/4k)	-	-	0.474	33.74	-	-
Random-Drop (4k/4k)	2k	500	0.500	33.39	0.504	33.48
	2k	1k	0.474	32.76	0.483	32.89
	4k	500	0.497	33.72	0.498	33.85
	4k	1k	0.506	33.40	0.518	33.48
Random-Drop (4.5k/4.5k)	2k	500	0.491	33.35	0.495	33.37
	4.5k	500	0.485	33.69	0.487	33.81
Baseline (6k/6k)	-	-	0.444	33.94	-	-
Random-Drop (6k/6k)	2k	500	0.468	33.46	0.471	33.46
	2k	1k	0.441	32.86	0.445	33.03
	6k	500	0.458	33.69	0.458	33.94
	6k	1k	0.473	33.60	0.472	33.71
Random-Drop (6.5k/6.5k)	2k	500	0.462	33.37	0.464	33.44
	6.5k	500	0.451	33.69	0.453	33.70

Tokenizer	N	k	Eff$_\alpha$	BLEU
Baseline (4k/4k)	-	-	0.474	33.74
Duplication (4k/4k)	100	3	0.594	32.37
	100	5	0.648	31.32
	500	3	0.583	32.26
	500	5	0.627	N/A
Baseline (6k/6k)	-	-	0.444	33.94
Duplication (6k/6k)	100	3	0.560	32.27
	100	5	0.612	31.60
	500	3	0.552	32.43
	500	5	0.598	30.57

Other Metrics

Tokenizer	N	k	PCT ↑	SEQ ↓	BLEU
Baseline (4k/4k)	-	-	0.461	25.50	33.74
Random-Drop (4k/4k)	2k	500	0.356	31.46	33.39
	2k	1k	0.233	40.37	32.76
	4k	500	0.405	29.23	33.72
	4k	1k	0.352	33.37	33.40
Random-Drop (4.5k/4.5k)	2k	500	0.356	31.46	33.35
	4.5k	500	0.402	27.93	33.69
Duplication (4k/4k)	100	3	0.590	25.50	32.37
	100	5	0.633	25.50	31.32
	500	3	0.571	25.50	32.26
	500	5	0.605	25.50	N/A

The Future

- Probably not too bad for intrinsic metrics
- Most natural tokenizers still follow the efficiency hypothesis
- Unclear about the relation to subword regularization
- Not yet tested on non-generation tasks (e.g., classification)
- Hopefully helps for designing better metrics!

† *Tokenization and the Noiseless Channel* (Zouhar et al., ACL 2023)

These research results were obtained partially from the commissioned research (No. 225) by National Institute of Information and Communications Technology (NICT), Japan.